Projected sales of main products in 2013
Distribution of marketing participation in the securities market
Revenue growth divisions
TYU division
FRT division
Share of market activity
Distribution of the securities market key players
FINAL SALES
SALES BY CATEGORY

BDA601 – BIG DATA AND ANALYTICS

Model Evaluation Infected Covid-19

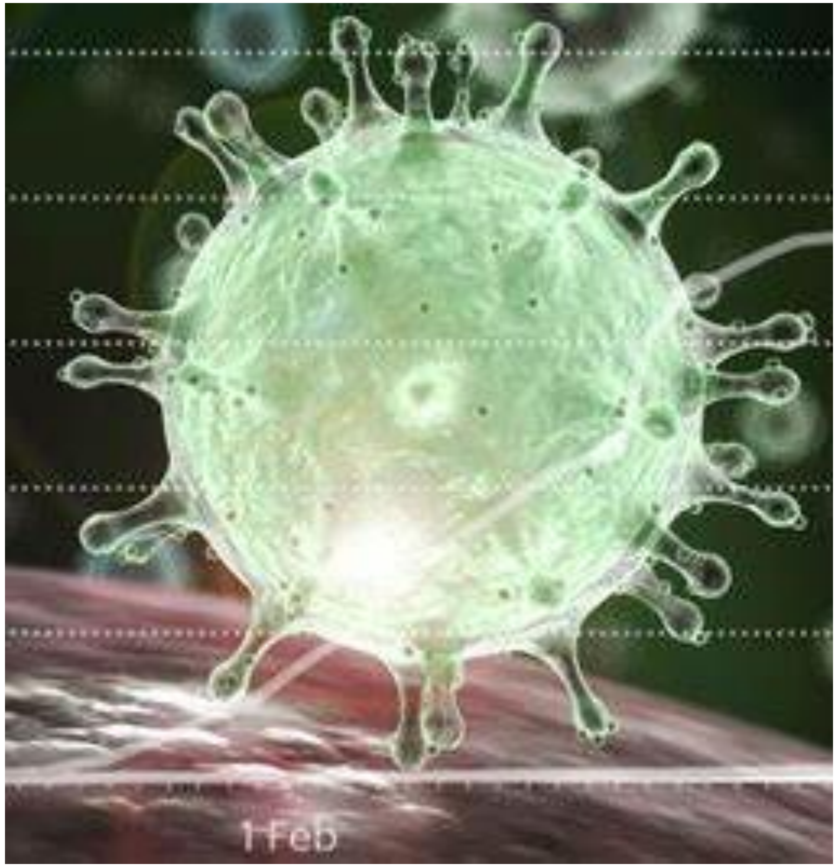
Beatriz Bernalte Gomez – A00075605

Introduction

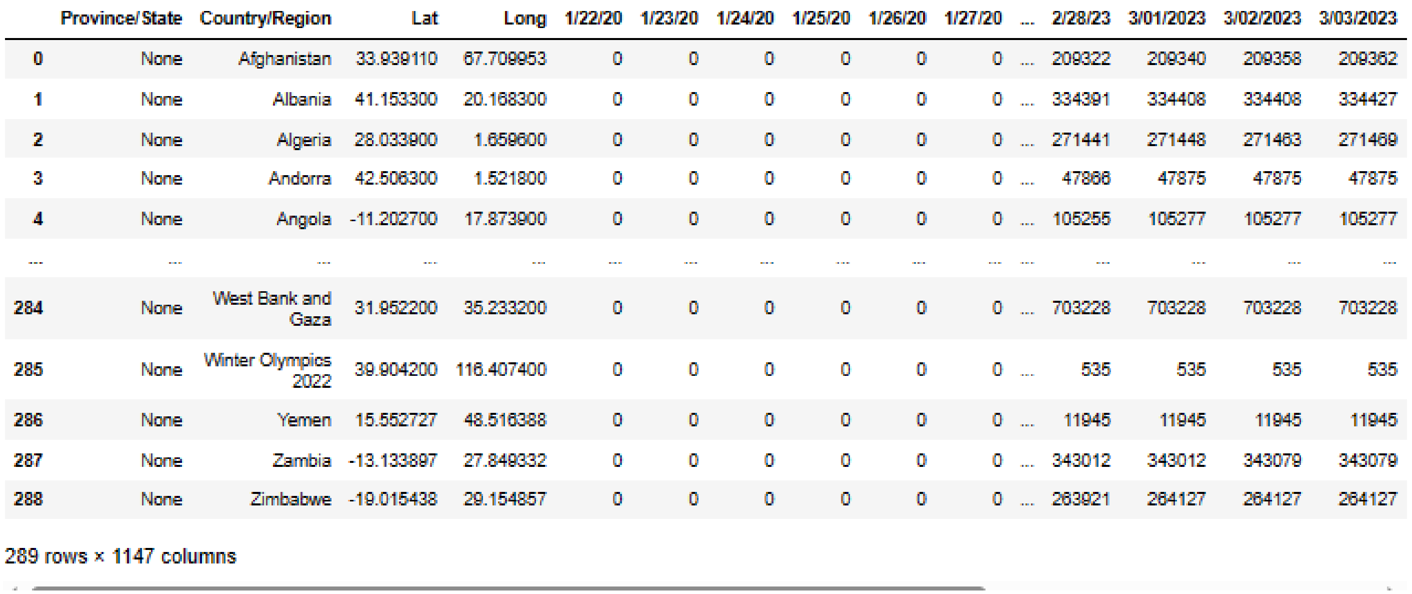
In recent times, the entire world has faced one of the greatest challenges of our generation the massive spread of a highly contagious virus which no one was prepared for.

During this period, big data analytics has become an invaluable tool for understanding, tracking, and controlling the disease.

In this presentation, we will explore how big data analytics has revolutionized our ability to monitor and respond to those infected with COVID-19, providing critical information for effective decision making.



Dataset



	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	...	2/28/23	3/01/2023	3/02/2023	3/03/2023
0	None	Afghanistan	33.939110	67.709953	0	0	0	0	0	0	...	209322	209340	209358	209362
1	None	Albania	41.153300	20.168300	0	0	0	0	0	0	...	334391	334408	334408	334427
2	None	Algeria	28.033900	1.659600	0	0	0	0	0	0	...	271441	271448	271463	271469
3	None	Andorra	42.506300	1.521800	0	0	0	0	0	0	...	47866	47875	47875	47875
4	None	Angola	-11.202700	17.873900	0	0	0	0	0	0	...	105255	105277	105277	105277
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
284	None	West Bank and Gaza	31.952200	35.233200	0	0	0	0	0	0	...	703228	703228	703228	703228
285	None	Winter Olympics 2022	39.904200	116.407400	0	0	0	0	0	0	...	535	535	535	535
286	None	Yemen	15.552727	48.516388	0	0	0	0	0	0	...	11945	11945	11945	11945
287	None	Zambia	-13.133897	27.849332	0	0	0	0	0	0	...	343012	343012	343079	343079
288	None	Zimbabwe	-19.015438	29.154857	0	0	0	0	0	0	...	263921	264127	264127	264127

289 rows × 1147 columns

- **Links dataset:**

Novel Coronavirus (COVID-19) Cases Data - Humanitarian Data Exchange (humdata.org)

- The dataset contains 289 rows and 1147 columns.
- The dataset contains information about countries/regions, their latitude and longitude, and the total count of COVID-19 dispositions per day from 01/22/2020 to 03/09/2023.

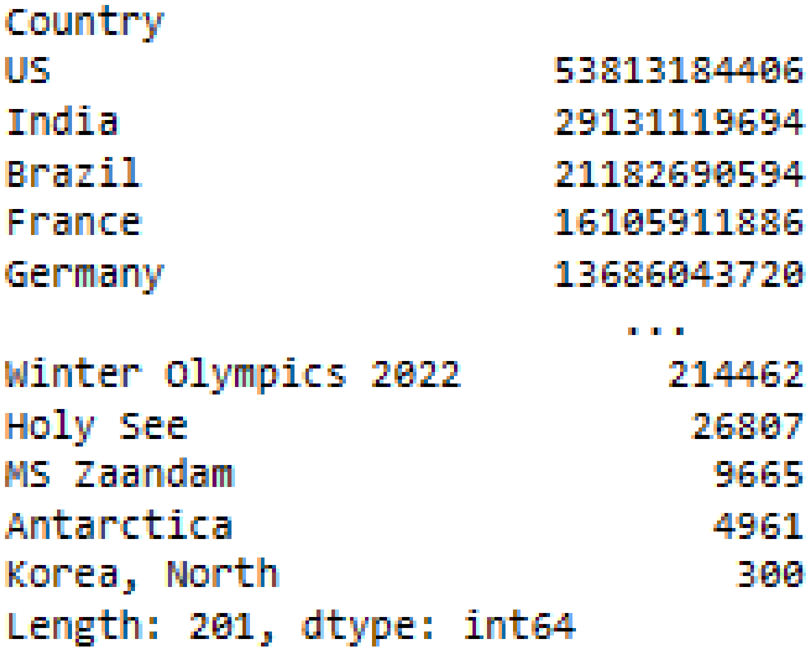
Analysis Methodologies

To carry out this analysis, Apache Spark methodologies have been used to prepare the data:

- Selection of columns with relevant data for the analysis.
- Management of missing values.
- Transform the data into the correct format.
- Group columns with dates in weeks.
- Unify the provinces and regions in their corresponding country.

Summary of countries

Total sum of infected by countries

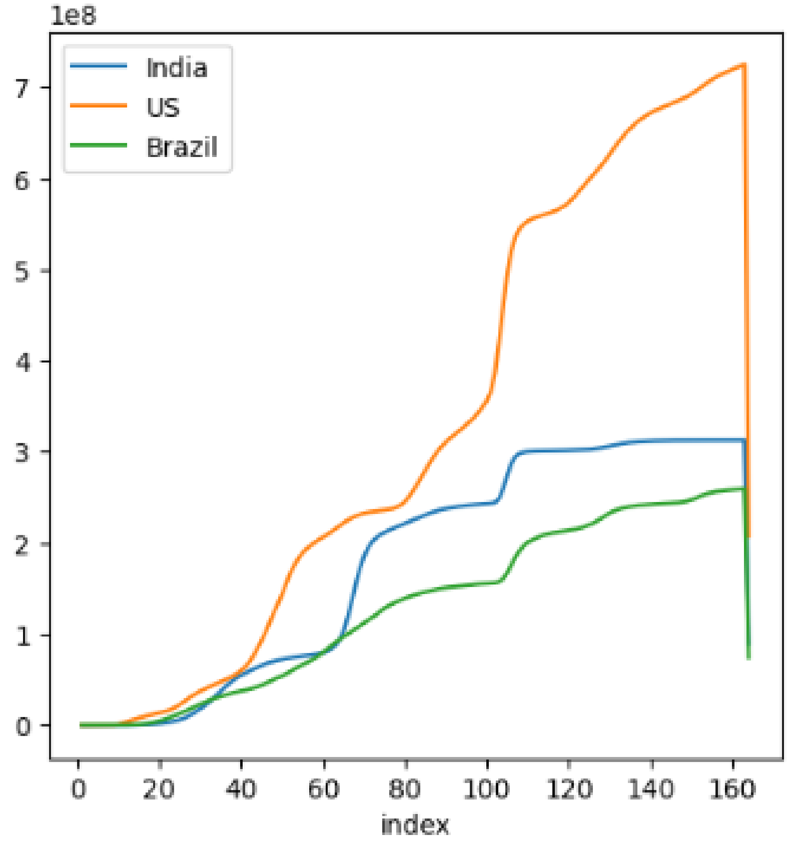


```
Country
US                     53813184406
India                  29131119694
Brazil                 21182690594
France                 16105911886
Germany                13686043720
                          ...
Winter Olympics 2022        214462
Holy See                     26807
MS Zaandam                    9665
Antarctica                    4961
Korea, North                   300
Length: 201, dtype: int64
```

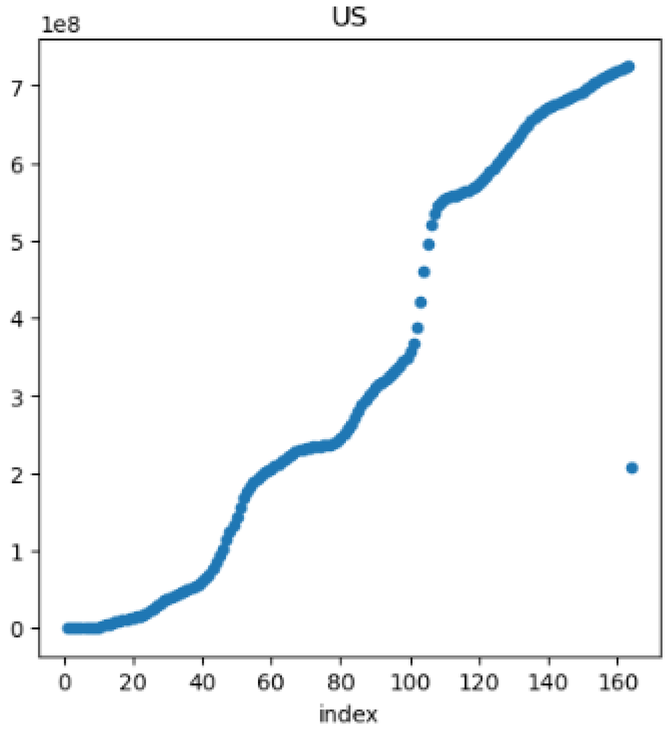
- Here we see the data of the total number of infected in 201 countries in the world.
- The countries with the most covid during this entire period have been the US, India and Brazil respectively.
- The least affected countries: Ms Zaandam, Antarctica and North Korea.

Top tree countries

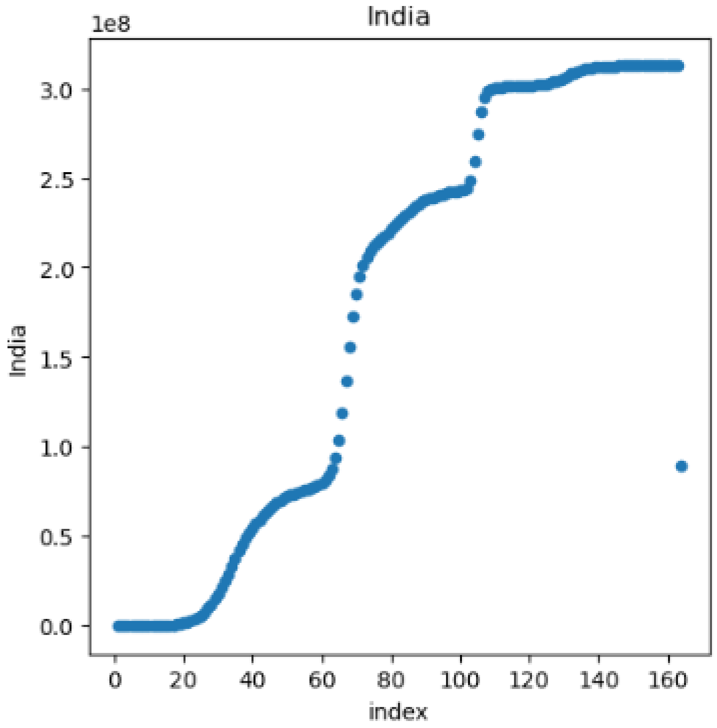
- The number of infected begins to rise from week 25.
- Week 40, US has a significant rise.
- During the time between weeks 75 to 100 the number of infected is maintained.
- Important rebound in week 100.
- As of week 110, India and Brazil remain without raising and lowering the number, on the other hand the trend of the US continues to rise continuously.



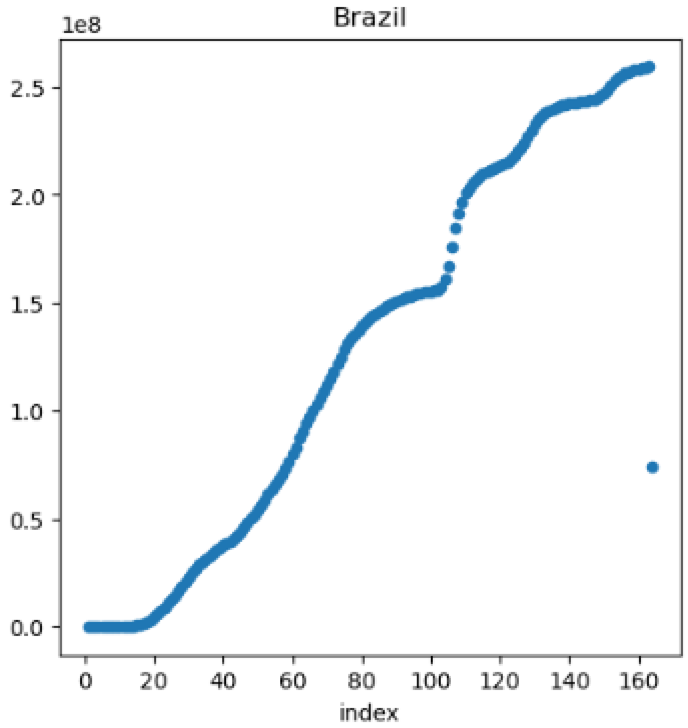
Linear Regression



R-squares: 0.965
Accuracy of 97%



R-square: 0.926
Accuracy of 93%

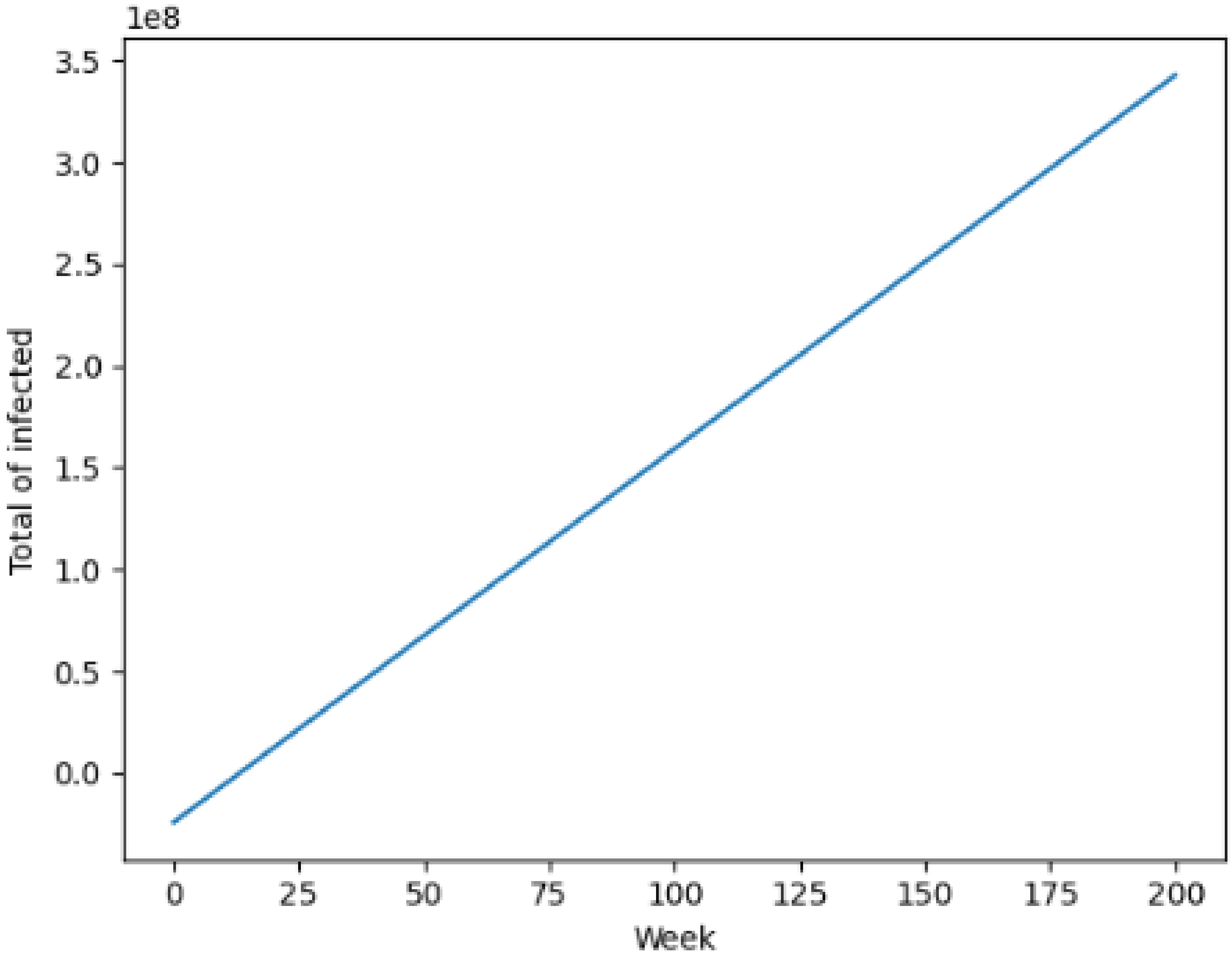


R-square: 0.977
Accuracy of 98%

Prediction

Model predictions:

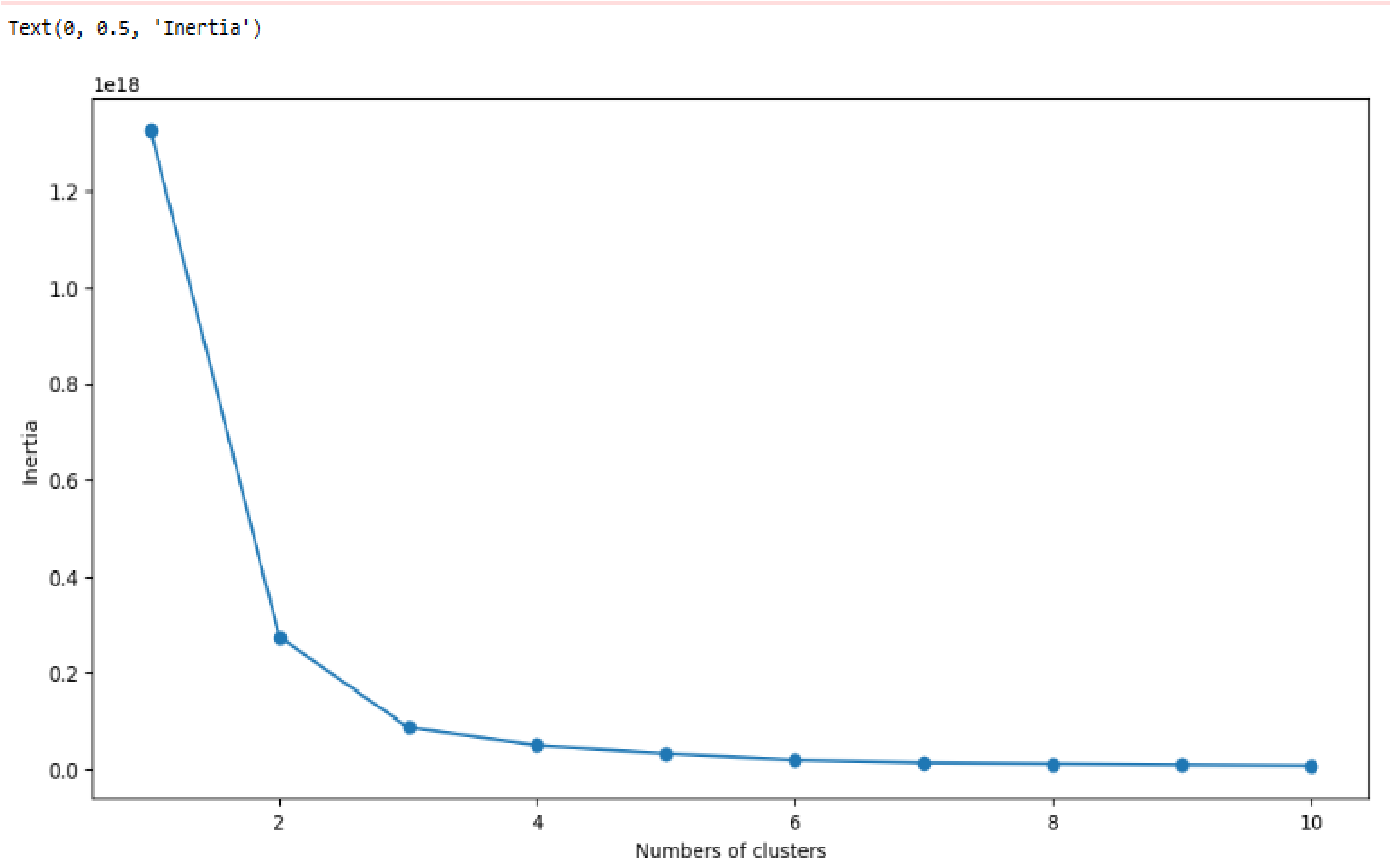
- Week 175 will be around 28 millions of infected.
- Week 200 Brazil would have 35 millions of infected.



Elbow

The objective of the elbow method is to find the number of clusters that best fit the data, minimizing the variance within each cluster.

As we can see, the point where the axis begins to stabilize is at 3, which means that we will adjust the model to 3 clusters.

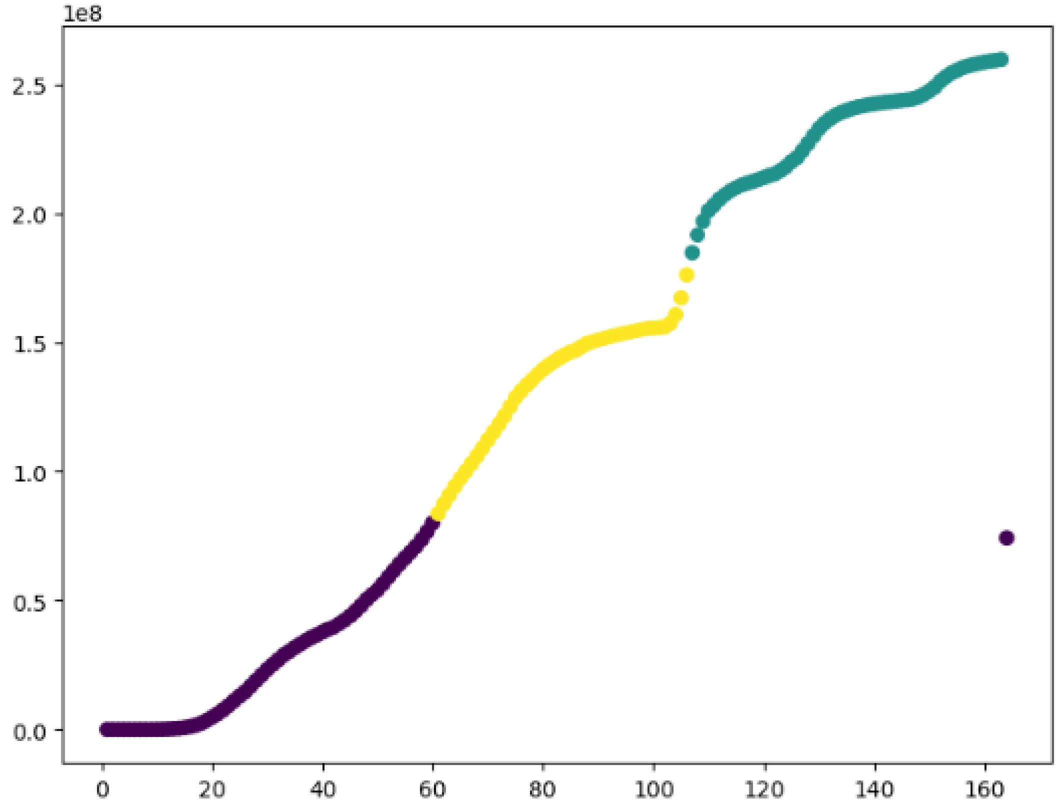


```
Text(0, 0.5, 'Inertia')
```

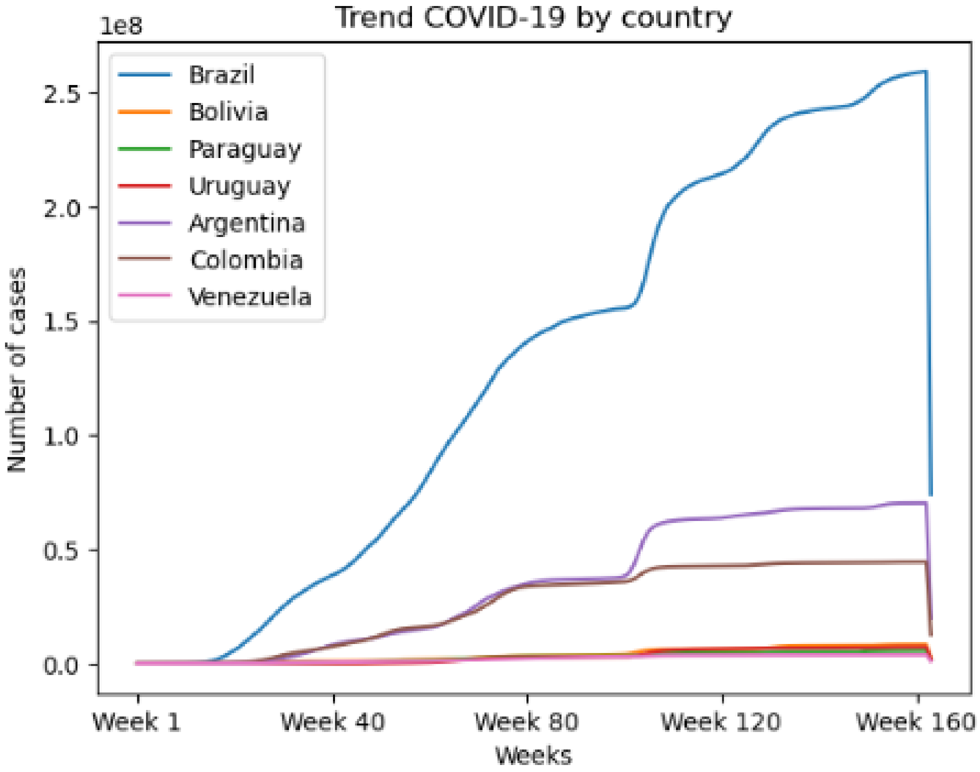
K-Means

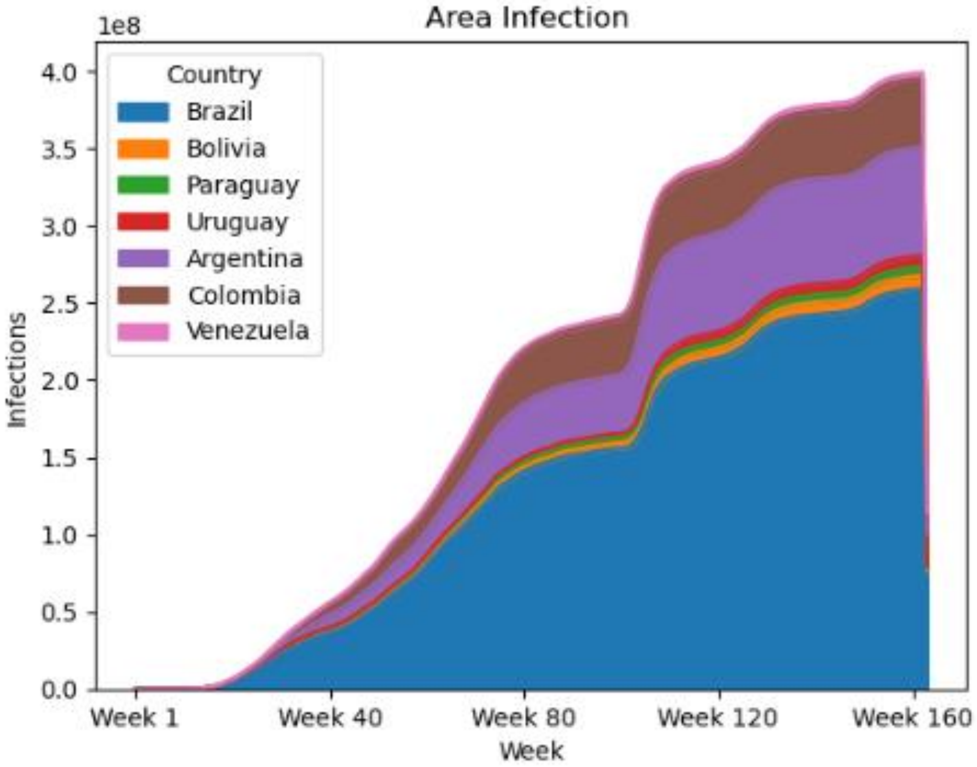
The model groups the data into 3:

- Group 1(purple), first stage of COVID with less than 10 million.
- Group 2 (yellow), second stage of COVID between weeks 60 and 110 with a total number of infected between 10 and 20 million.
- Group 3 (green), third stage with more than 20 million infected.



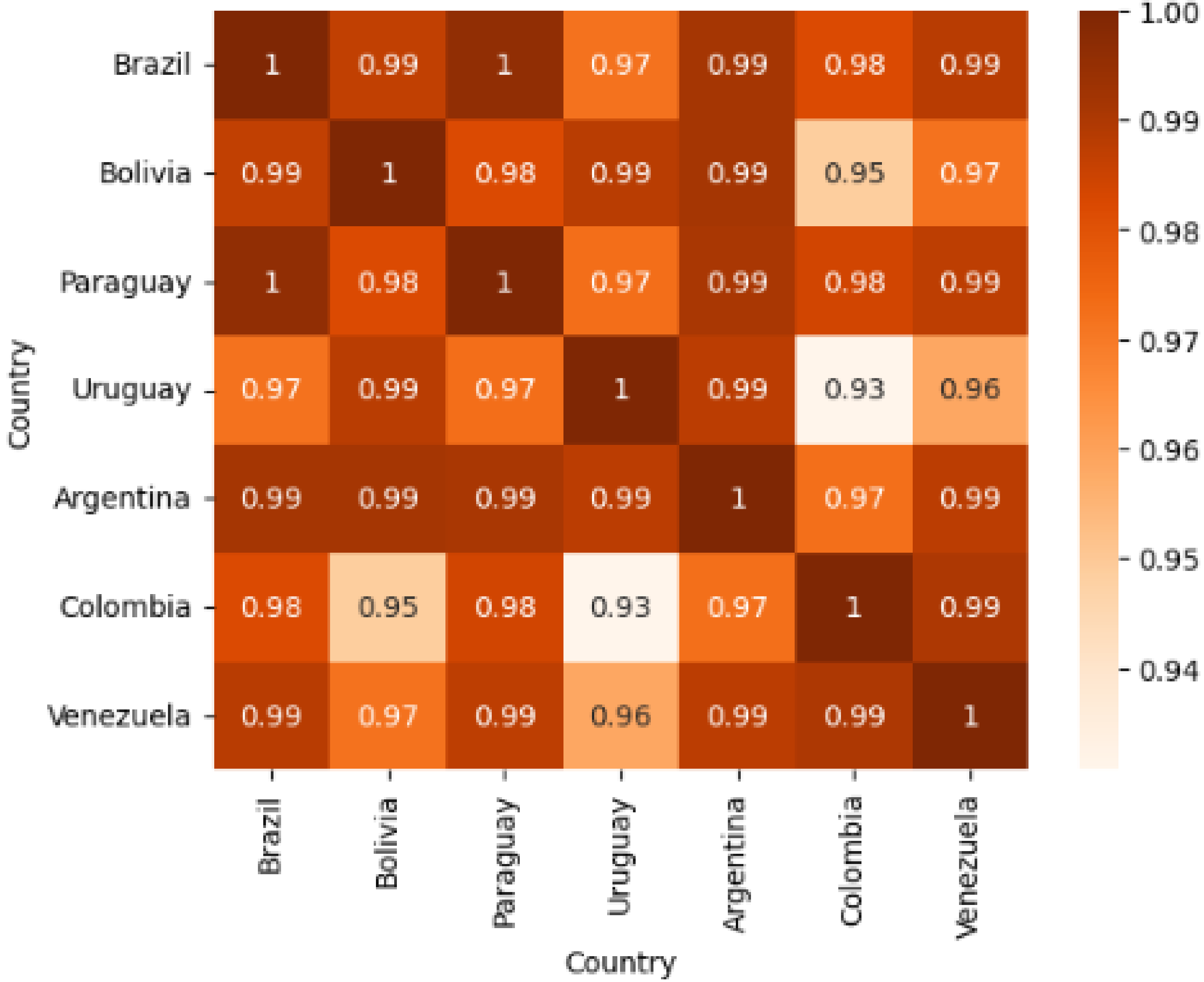
Neighbor Countries





Countries’s Correlation

- In this correlation graph, all neighboring countries have a high correlation with Brazil, especially Paraguay, which indicates that the infection patterns are similar.



Conclusion

In conclusion, it is critical that neighboring countries take immediate action to control the spread of COVID-19, as the infection rate continues to rise steadily. The COVID-19 pandemic has proven to be highly contagious and can have devastating consequences for a nation's public health and economy.

In addition, it is recommended to closely monitor the evolution of the pandemic and adapt the Machine Learning models to obtain more patterns.